Improving Human Pose Recognition Accuracy Using CRF Modeling

Vivek Sharma[◇⋆], Frank Dittrich[†], Şule Yildirim-Yayilgan[⋆], and Luc Van Gool[◇∓]
[◇]KU Leuven, ESAT-PSI, iMinds [⋆]Gjøvik University College [†]KIT [∓]ETH Zürich
{vivek.sharma,luc.vangool}@esat.kuleuven.be sule.yayilgan@hig.no frank.dittrich@kit.edu

1. Introduction

Interest in robotics in the domain of manufacturing industry has shown an outstanding growth recently in scenarios where human beings and robots are present simultaneously. Humans and robots often share the same workspace and this poses a lot of threats to the human safety issues [1] *e.g*. in manufacturing industry, in automobile industry where automobile components are integrated, in medical industry where minimally-invasive-surgery is facilitated and so on. In the proposed approach, segmentation is defined as a classification task and is used for pixelwise object class labeling of human body-parts. Depth measurements from a KINECT RGB-D ceiling sensor are obtained in order to do the pixelwise object class labeling. The ultimate intended use is in the safe human-robot collaboration (SHRC) and interaction (SHRI) domains for challenging domestic and industrial environments. Within this scope, a pairwise conditional random field (CRF) approach is used for labeling. CRF is formulated in terms of an energy minimization (EM) problem while an efficient random decision forest (RDF) is used for classification. In [4], we show how an RDF classifier is used for pixelwise classification of human body-parts using deph data. We found that there exists misclassification of labels assigned to each pixel and that should be minimized for feasible and practical human-robot cooperation. This work builds on top of our previous work Dittrich *et al*. [4] in order to improve recognizing human body-parts.

2. Proposed Approach

In energy minimization (EM) based labeling problems [3, 6], one aims to assign a label to each pixel which minimizes the energy and gives the most optimal labeling. The EM or CRF energy is defined as:

$$E(\mathbf{x}) = E_{data}(\mathbf{x}) + E_{smooth}(\mathbf{x}) \quad (1)$$

where $\mathbf{x}$ being an arbitrary configuration of assigning each pixel a label which is both pairwise regularized smooth [2] and consistent with the observed depth data.

In our approach, the unary ($E_{data}(\mathbf{x})$) term is the likelihood of an object label assigned to pixel, obtained from the RDF classifier [4]. The pairwise prior ($E_{smooth}(\mathbf{x})$) term encodes a smoothness prior and takes the form of classical Ising-Potts model [2], which can be efficiently minimized by using α-expansion [3]. The α-expansion algorithm finds the optimal subset of pixels that must switch from an arbitrary label to a fixed label (α) which minimizes the energy function. The function converges to the global optimal solution when no lower energy solution can be found.

3. Data Collection

A dataset of pixelwise RGB-D data of human body-parts (*head, body, upper-arm, lower-arm, hand and legs*) has been generated synthetically in a virtual environment [5] using a KINECT sensor. KINECT skeleton estimations [4] reduce the computational expense in comparison with the motion capture technique [7]. The synthetic dataset composes of highly variable combinations of human poses and shapes (*e.g. sitting, standing, walking, working, dancing, swinging, boxing, tilting, bending, bowing, and stretching*) in a mixed way with angled single and both arms and many more. The human height ranges between 160-190 cm. Although we train the RDF classifier using the synthetic dataset, we use real-world human data for testing.

4. Results and Conclusion

Table.1 presents the results obtained using our proposed approach and evaluations in terms of average Precision (mAP) and Recall (mAR) are given. It is observed that CRF modeling on top of using an RDF classifer versus using an RDF Classifier only [4] improves the segmentation performance by approximately $\mathbf{5.6}\%$ in mAR and $\mathbf{9.9}\%$ in mAP. The improvement of performance for human body-part segmentation is more meaningful for SHRC and SHRI domain. In Fig.2, we show the segmentation results for different human pose and shape configurations with varying human height.

In [7], Shotton *et al*. use a number of training frames (F) 300K/tree where 2000 depth values (PC) were extracted

	Avg	Head	Body	UArm	LArm	Hand	Legs
RDF_{mAR}	**0.787**	0.931	0.795	0.718	0.612	0.699	0.972
RDF_{mAP}	**0.651**	0.971	0.632	0.718	0.709	0.639	0.238
Ours$-CRF_{mAR}$	**0.843**	0.946	0.835	0.849	0.651	0.791	0.987
Ours$-CRF_{mAP}$	**0.750**	0.975	0.849	0.741	0.777	0.802	0.361

Table 1. Confusion matrix based mAR and mAP measures for RDF and CRF modeling using real-world data test

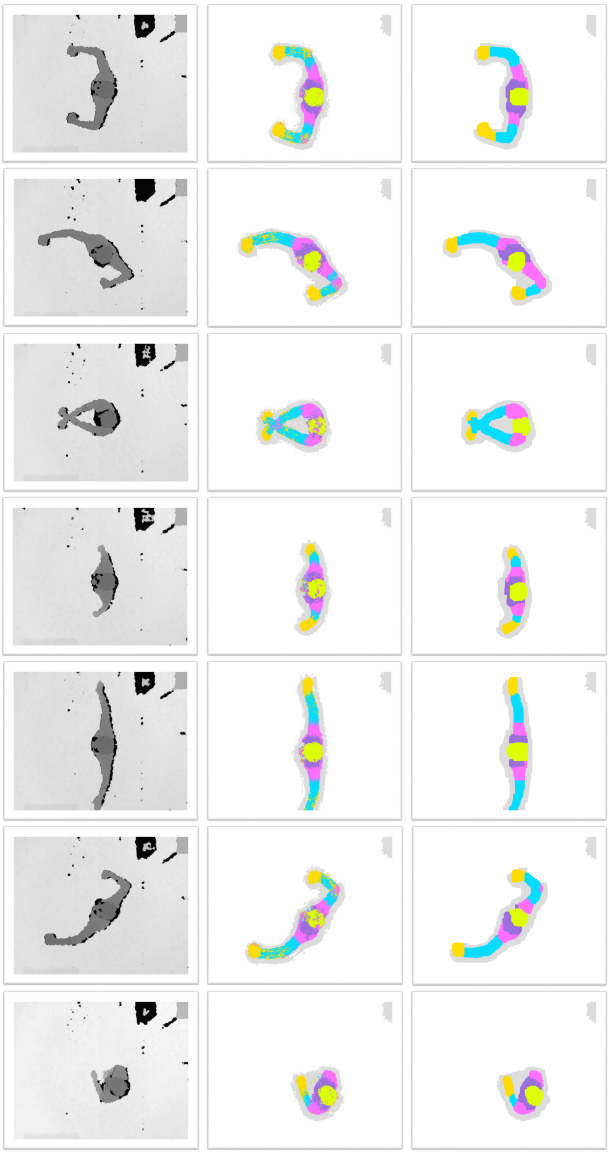

Figure 1. Segmentation results based on real-world test depth data. The first column shows the test data, the second and third columns show the predictions obtained from RDF and CRF modeling.

for each of 31 body-parts of human which took approximately a day for training a decision tree on a 1000 core cluster. Shotton *et al*. approach is computationally very expensive and consumes a large memory. While in our case F=1600/tree where PC=300 are extracted for each of our defined 6 body-parts of human is sufficient enough for producing almost comparable results. The training time of RDF with our highly optimized parameter set takes 43 minutes. Hence reducing computational expense and memory consumption. Also our work can distinguish subtle changes such as crossed-arms which is not possible in [7].

5. Comparison with the State-Of-The-Art

As a baseline with [4], we compare our performance results using "*top-view*" as a comparison parameter for human body-parts classification. Fig.2 shows comparison of the per-joint proposals of the human body-parts classification. Our results for per-joint classification of human body-parts shows a significant improvement over [4].

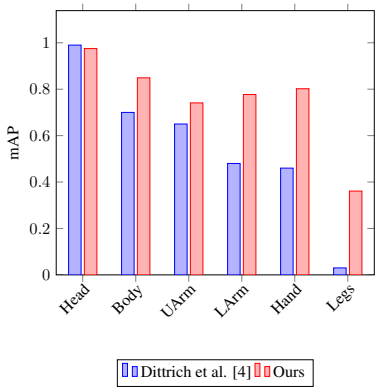

Figure 2. Comparison with [4]. Our approach is sufficient for producing almost comparable and better results for localizing the joints of the human body-parts.

Ackowledgements: This work is supported by the BMBF funded project AMIKA and the EU project ROVINA.

References

[1] *Fraunhofer IFF*, 2014.

[2] Y. Boykov and G. Funka-Lea. Graph cuts and efficient n-d image segmentation. *IJCV*, 2006.

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. PAMI*, 2001.

[4] F. Dittrich, V. Sharma, H. Wörn, and S. Yayilgan. Pixelwise object class segmentation based on synthetic data using an optimized training strategy. In *ICNSC*, 2014.

[5] M. Freese, S. P. N. Singh, F. Ozaki, and N. Matsuhira. Virtual robot experimentation platform v-rep: A versatile 3d robot simulator. In *SIMPAR*, 2010.

[6] V. Sharma, F. Dittrich, S. Yayilgan, and L. V. Gool. Efficient real-time pixelwise object class labeling for safe human-robot collaboration in industrial domain (accepted). In *ICML Workshops*, 2015.

[7] J. Shotton, R. B. Girshick, A. W. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake. Efficient human pose estimation from single depth images. *IEEE Trans. PAMI*, 2013.